Clustering-Enhanced Stochastic Gradient MCMC for Hidden Markov Models

by

Rihui Ou

Department of Statistical Science
Duke University

Date: ______________________________
Approved:

____
David Dunson, Supervisor

____
Sayan Mukherjee

____
Galen Reeves

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2019

ABSTRACT

Clustering-Enhanced Stochastic Gradient MCMC for Hidden Markov Models

by

Rihui Ou

Department of Statistical Science
Duke University

Date: ______________________________
Approved:

____
David Dunson, Supervisor

____
Sayan Mukherjee

____
Galen Reeves

An abstract of a thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2019

Abstract

MCMC algorithms for hidden Markov models, which often rely on the forward-backward sampler, suffer with large sample size due to the temporal dependence inherent in the data. Recently, a number of approaches have been developed for posterior inference which make use of the mixing of the hidden Markov process to approximate the full posterior by using small chunks of the data. However, in the presence of imbalanced data resulting from rare latent states, the proposed minibatch estimates will often exclude rare state data resulting in poor inference of the associated emission parameters and inaccurate prediction or detection of rare events. Here, we propose to use a preliminary clustering to over-sample the rare clusters and reduce variance in gradient estimation within Stochastic Gradient MCMC. We demonstrate very substantial gains in predictive and inferential accuracy on real and synthetic examples.

Acknowledgements

I would like to express my special gratitude to my supervisor Dr. David Dunson, for his invaluable advice and inspiration as a mentor during this project. I would not have finished this thesis without his help. It is my privilege to work with him. I would like to thank my collaborator Dr. Alexander Young, for his insightful advice on numerical experiments and inspirational discussion. I would like to thank for Dr. Sayan Mukherjee and Dr. Galen Reeves for serving as my committee members. I want to thank Dr. Merlise Clyde, Dr. Alexander Volfovsky and Dr. Robert Wolpert who introduced me into the world of Bayesian statistics. I also want to thank Dr. Surya Tokdar for his invaluable advice on research and encouragement. Words cannot express my appreciation to my parents for their support during my life. I would also like to thank all of my friends.

Contents

List of Figures

Chapter 1

Introduction

As time series data become increasingly large, traditional algorithms for inference suffer from the added computational burden and are often intractable. As a result, there has been an increased interest in the development of scalable algorithms. We are particularly interested in obtaining reliable uncertainty quantification (UQ) in inferences and prediction. There is a recent literature on relevant methods, including approaches that subsample 'chunks' of the times series [Joh14, HSS15] and data thinning [JW16]. In the present work, we focus on Hidden Markov Models (HMMs) – well-regarded for their versatility in characterizing a breadth of phenomena including protein folding dynamics [SZG+11], credit ratings [PCX16], speech recognition [HAJ90], stock market forecasting [HN05] and rare event detection [ZGBM05, CHSP04]. In particular, we consider the case where the time series exhibits normal behavior over large intervals of time interrupted intermittently by different transient dynamics associated with one or more rare states in the latent Markov chain akin to the dynamics of network attacks [OMSH03] and solar flares [HW13].

Due to the temporal dependence inherent in sequential data, methods for inferring HMM parameters – such as Monte Carlo methods [Sco02], Expectation-Maximization [Bis06], and variational Bayes' [JW14] – often rely on an implementation of the forward-backward algorithm (FB) named as such due to the need to make both a forward and backward pass through the entire time series. At each iteration, one attains a local update of the unknown latent states in the HMM using FB to obtain marginal distributions of these states. This is followed by a global update of the parameters for the emission distributions. As the FB must pass through the entire sequence of observations in the time series, it results in a time complexity which is linear in the number of samples, burdensome for very large time series.

One natural approach to addressing the computational burdens is to subsample the data. Unfortunately, while a number of subsampling strategies have been developed in the *iid* setting [CB17, SBB$^+$16, HCB16, QKVT18, FZ17], the sequential nature of time series makes this strategy more difficult. Most relevant to the present work, [FXLF14] constructed a Stochastic Variational Inference HMM (SVI-HMM) algorithm which breaks long sequences apart to construct minibatch samples. The approach is justifiable due to the inherent memory decay in HMMs when the time series is long. Unfortunately, as noted in [MFF17], this algorithm suffers from a number of drawbacks, in particular, the well known under-estimation of posterior correlation and limitations to conjugate emission distribution. Through an alternative formulation of the HMM, [MFF17] also made use of the underlying mixing of the latent states to separate the time series to develop a Stochastic Gradient MCMC (SG-MCMC) framework for inferring the emission parameters and transition probabilities. Their approach demonstrated drastically improved computational performance compared to traditional FB. In this SG-MCMC algorithm, estimates of the correlation decay rate of the underlying Markov chain are obtained to determine appropriately long buffers separating subseries into effectively independent chunks. These subseries of the full data are then subsampled to obtain minibatch estimates of the full data log-posterior gradient.

There is one large drawback common to both SVI-HMM and SG-MCMC. In the presence of one or more rare states, minibatches will often fail to include any portion of the time series in which transient dynamics occur. As a result, one should expect current minibatch based estimation to fail to discover rare state(s), even though these states are often of particular interest. Motivated by this observation, we propose to first cluster the subseries, resulting in a biased subsampling strategy which has a high probability of including rare dynamics, which we then implement within SG-MCMC. Our Cluster-enhanced Stochastic Gradient MCMC (CSG-MCMC) approach adds minimal computational complexity while drastically improving the accuracy of gradient estimates. We provide numerical evidence from synthetic data experiments that the clustering based estimates have smaller variance than those of [MFF17] thereby minimizing error and improving predictive performance. Additional experiments are provided using heart rate data from the MIMIC-III database [JPS+16].

This work is done in collaboration with my advisor, Dr. David Dunson, and Dr. Alexander Young.

Chapter 2

Background

2.1 Hidden Markov Models and Motivation

HMMs consist of discrete-valued latent states $x_t \in \{1, \dots, K\}$ generated by a Markov chain and the corresponding observations y_t generated by an emission distribution determined by the state x_t. Specifically, the joint distribution of $\mathbf{y} = (y_1, \dots, y_T)$ and $\mathbf{x} = (x_0, \dots, x_T)$ factorizes as

$$p(\mathbf{x}, \mathbf{y} \mid A, \phi) = p(x_0) \prod_{t=1}^{T} p(y_t \mid x_t, \phi) p(x_t \mid x_{t-1}, A) \tag{2.1}$$

where $A_{i,j} = P(x_t = i \mid x_{t-1} = j)$ is the Markov transition matrix for the latent state, $\phi = \{\phi_k\}_{k=1}^{K}$ consists of parameters of the associated emission distributions, and $p(x_0)$ is the distribution of the initial state. Throughout, we assume the latent Markov chain is recurrent and irreducible and that the latent states are at stationarity so that x_0 is drawn from the unique stationary distribution for A.

As a motivating example, consider the following extreme case. Suppose there are two latent states with transition matrix

$$A = \begin{bmatrix} 1 - \epsilon & 0.1 \\ \epsilon & 0.9 \end{bmatrix} \tag{2.2}$$

which has stationary distribution $\pi = \frac{1}{10\epsilon+1}(1, 10\epsilon)^T$. Furthermore, consider Gaussian emissions

$$(y_t \mid x_t = j) \sim \mathcal{N}\big(\mu_j, 10^{-4}\big),\ j = 1, 2. \tag{2.3}$$

where $\mu_j = j - 1$. A realization of this process, shown in Figure 2.1 with $\epsilon = 10^{-4}$, shows rare visits to a neighborhood of one.

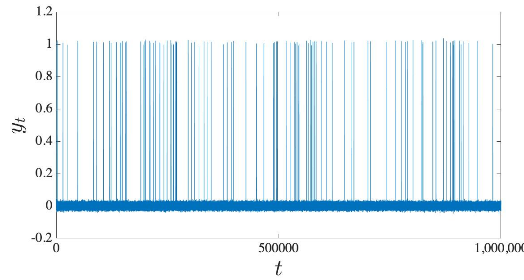

Figure 2.1: One realization of the HMM with transition matrix (2.2) and emission distributions (2.3) is shown above.

Now, assume that only the common variance, $\sigma^2 = 10^{-4}$, of the emission distributions is known and we are interested in inference on the transition matrix and means. We use the conjugate priors

$$\mu_j \sim \mathcal{N}(\mu_j^{(0)}, \sigma_j^{2(0)}), j = 1, 2$$

$$\mathbf{a}_j \sim \text{Dirichlet}(\mathbf{a}_j^{(0)}), j = 1, 2$$

where $\mathbf{a}_j$ corresponds to column j of A. Given the minimal overlap of the two emission distributions the latent states can be (almost exactly) identified through the observations. Thus, we treat $\mathbf{x}$ as known given $\mathbf{y}$ thereby ignoring the sequential dependence in the data. In this case, the posterior distributions are

$$\begin{aligned} \mu_j \mid \mathbf{y} &\sim \mathcal{N}\left(\frac{1}{\sigma_j^{2(N_i)}}\left(\frac{\mu_j^{(0)}}{\sigma_j^{2(0)}} + \frac{N_i \bar{Y}_i}{\sigma^2}\right), \sigma_j^{2(N_i)}\right) \\ \mathbf{a}_j \mid \mathbf{y} &\sim \text{Dirichlet}(\mathbf{a}_j + (N_{j1}, N_{j2})^T) \end{aligned} \tag{2.4}$$

where N_j is the number of observations in state j, $\bar{Y}_j$ is the sample mean of the observations in state j, N_{jk} is the number of transitions from state j to state k, and $\sigma_j^{2(N_i)} = \frac{\sigma^2 \sigma_j^{2(0)}}{\sigma^2 + N_j \sigma_j^{2(0)}}$ is the posterior variance of μ_j.

Suppose now that we subdivide the full time series into T/L subchains each of length L and construct a posterior sampling algorithm as follows. At each step, we draw uniformly at random $S \ll T/L$ subchains and sample from (2.4) using only these subsampled chains. When $0 < \epsilon \ll 1$, state 2 will be so rare that we can expect

$$N_2 = \bar{Y}_2 = N_{21} = N_{22} = 0$$

at (almost) every iteration so that the posterior estimates of μ_2 and $\mathbf{a}_2$ will be essentially equivalent to their priors.

Of course, poor inference for the rare state parameters may also occur using the full data. However, one should expect this minibatch subsampling approach to greatly exacerbate the issue particularly when the subchains cover a small proportion of the full chain, i.e. $SL \ll T$. As both the SVI-HMM of [FXLF14] and SG-MCMC of [MFF17] are based on this sampling approach, it is reasonable to expect both methods will suffer when there are one or more rare latent states. In fact, it is possible that such an approach will fail to detect rare states all together.

Consider the alternate posterior sampling algorithm which avoids this issue. First, we cluster the subchains into groups C_1, which contains subchains where all latent states are one, and C_2, which contains subchains with at least one observation from state two. We then use a stratified subsample in (2.4) by drawing S_j subsamples uniformly at random from cluster C_j where $S_1 + S_2 = S$. At each step, we can be assured that N_2, Y_2, N_{21}, and N_{22} will be positive thereby improving the inference regarding the rare state. We investigate this subsampling strategy in concert with SG-MCMC which we now review.

2.2 Stochastic Gradient MCMC for HMMs

Consider the alternate formulation of the HMM which marginalizes out the hidden states to attain the marginal likelihood,

$$p(\mathbf{y} \mid \theta) = \mathbf{1}^T \left(\prod_{t=1}^{T} P(y_t) A \right) \pi \tag{2.5}$$

where $P(y_t)$ is the diagonal matrix with entries $P_{jj}(y_t) = p(y_t \mid x_t = j, \phi)$, $\mathbf{1}$ is a K-dimensional vector of ones, and $\pi = A\pi$ is the stationary distribution for A. Given a prior $p(\theta)$, the posterior distribution of θ given observations $\mathbf{y}$ is then,

$$p(\theta \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta)p(\theta)}{p(\mathbf{y})} \propto p(\mathbf{y} \mid \theta)p(\theta) \tag{2.6}$$

Assuming a continuous parameter θ, marginalizing the latent states allows one (in theory) to make use of sampling algorithms such as Hamiltonian Monte Carlo or SG-MCMC. Unfortunately, both methods require calculation of the gradient of $U(\theta) = -\log p(\theta \mid \mathbf{y})$ with components

$$\frac{\partial}{\partial \theta_j} U(\theta) = -\frac{\sum_{t=1}^{T} \mathbf{q}_{t+1}^T \left[\frac{\partial}{\partial \theta_j} (P(y_t)A) \right] \pi_{t-1}}{\mathbf{1}^T \left(\prod_{t=1}^{T} P(y_t)A \right) \pi^T} - \frac{\partial}{\partial \theta_j} \log p(\theta) \tag{2.7}$$

where

$$\mathbf{q}_{t+1}^T = \mathbf{1}^T \prod_{k=t+1}^{T} [P(y_t)A], \quad \pi_{t-1} = \left(\prod_{k=1}^{t-1} P(y_t)A \right) \pi. \tag{2.8}$$

When T is large, evaluating the gradient becomes intractable largely due to the repeated matrix multiplication needed to calculate $\mathbf{q}_{t+1}^T$ and π_{t-1}.

To circumvent this issue, [MFF17] construct gradient approximations utilizing minibatch subsampling in conjunction with the mixing of the latent Markov chain. We summarize their main idea in a slightly simpler construction.

Fix $L \ll T$ with L odd. For $\tau \in \{(L-1)/2, \ldots, T-(L-1)/2\}$, consider the subchain $\mathbf{y}_\tau = (y_{\tau-(L-1)/2}, \ldots, y_\tau, \ldots, y_{\tau+(L-1)/2})$ of length L centered at $t = \tau$ and let

$$P(\mathbf{y}_\tau) = \prod_{t=\tau-(L-1)/2}^{\tau+(L-1)/2} P(y_t)A.$$

Now partition the full time series into the T/L sequential subchains such that

$$\mathbf{y} = \{\mathbf{y}_{\tau_1}, \ldots, \mathbf{y}_{\tau_{\frac{T}{L}}}\}$$

Here, we fix the subchain centers $\tau_k = \frac{(2k-1)L+1}{2}$.

We may then express (2.7) using these T/L subseries as

$$\frac{\partial}{\partial \theta_j} U(\theta) = -\sum_{k=1}^{T/L} \frac{\mathbf{q}^T_{\tau_k+(L+1)/2} \frac{\partial P(\mathbf{y}_{\tau_k})}{\partial \theta_j} \pi_{\tau_k-(L+1)/2}}{\mathbf{q}^T_{\tau_k+(L+1)/2} P(\mathbf{y}_{\tau_k}) \pi_{\tau_k-(L+1)/2}} - \frac{\partial \log[p(\theta)]}{\partial \theta_j}. \tag{2.9}$$

Similar to (2.7), the calculation of $\mathbf{q}^T_{\tau_k+(L+1)/2}$ and $\pi_{\tau_k-(L+1)/2}$ will be computationally demanding when T is large as they must pass over the full series after and before $\mathbf{y}_{\tau_k}$ respectively. As such, it is desirable to find approximations. Assume A is known and let B be a buffer length longer than the inverse of the spectral gap of A. Given the memory decay in the latent states, any data occurring more than B steps before or after $\mathbf{y}_\tau$ will be essentially independent of it. Thus, we may make the following approximations.

$$\begin{aligned} \mathbf{q}^T_{\tau_k+(L+1)/2} &\approx \bar{\mathbf{q}}^T_{\tau_k+(L+1)/2} = \mathbf{1}^T \prod_{t=\tau_k+(L+1)/2}^{\tau_k+(L+1)/2+B} P(y_t)A \\ \pi_{\tau_k-(L+1)/2} &\approx \bar{\pi}_{\tau_k-(L+1)/2} \\ &= \left(\prod_{t=\tau_k-(L+1)/2-B}^{\tau_k-(L+1)/2} P(y_t)A \right) \pi \end{aligned} \tag{2.10}$$

where π is the invariant distribution of A. These approximations are simply truncations of the products in (2.8) making them much less demanding to calculate.

A second approximation arises by using a random subsample of S subchains in the calculation of (2.9). The approximated gradient $\widetilde{\nabla U(\theta)}$ has components

$$\widetilde{\frac{\partial}{\partial \theta_j} U(\theta)} = -\frac{T}{LS} \sum_{k \in \mathcal{S}} \frac{\bar{\mathbf{q}}^T_{\tau_k+(L+1)/2} \frac{\partial P(\mathbf{y}_{\tau_k})}{\partial \theta_j} \bar{\pi}_{\tau_k-(L+1)/2}}{\bar{\mathbf{q}}^T_{\tau_k+(L+1)/2} P(\mathbf{y}_{\tau_k}) \bar{\pi}_{\tau_k-(L+1)/2}}$$
$$- \frac{\partial \log p(\theta)}{\partial \theta_j} \tag{2.11}$$

where $S = |\mathcal{S}|$ and $\mathcal{S}$ is a subset of the subseries centers, $\{\tau_k\}_{k=1}^{T/L}$. Taking $S \ll T/L$ futher reduces the computational burden of approximating $\nabla U(\theta)$.

To ensure independence of the terms within (2.11) some care must be made to ensure each subsampled chunk and the B observations before and after it are sufficiently separated. This is equivalent to the requirement that

$$\bigcap_{j,k} \{\tau_k - L - B - \nu, \ldots \tau_k + L + B + \nu\} = \emptyset$$

for $j, k \in \mathcal{S}$ with $j \neq k$ where ν is a minimum gap between subseries. [MFF17] use a more flexible sampling approach which enforces this gap condition without fixing the subseries centers *a priori*. However, we do note that when $SL \ll T$, poor coverage of rare latent states should be expected.

The SG-MCMC algorithm proceeds using the gradient estimates in (2.11). Of course, in practice the transition matrix A is unknown, and therefore so are its invariant distribution and spectral gap. The authors address this issue naturally by using updated estimates of A to estimate B and π. See [MFF17] for the full details including technical details regarding sampling from the simplex in estimating the columns of A. A more general discussion of SG-MCMC can be found in [MCF15]. For now, we summarize the SG-MCMC method in Algorithm 1.

Algorithm 1 SG-MCMC for HMMs [MFF17]

set N_{step} = number of SG steps per iteration
initialize $A^{(0)}$ and $\phi^{(0)}$
  for $n = 1, 2, \ldots, N$ **do**
    Estimate the spectral gap and buffer length using $A^{(n)}$
    Sample subchains S of length L.
    for $s = 1, \ldots, N_{\text{steps}}$ **do**
      Update $\hat{A}^{(s)}$
    end for
    Calculate $\hat{A}^{(s)} = \frac{1}{N_{\text{steps}}} \sum_{s=1}^{N_{\text{steps}}} \hat{A}^{(s)}$
    Set $A_{i,j} \leftarrow |\hat{A}_{i,j}| / \sum_i |\hat{A}_{i,j}|$
    for $s = 1, \ldots, N_{\text{steps}}$ **do**
      Update $\phi^{(s)}$
    end for
    Set $\phi = \frac{1}{N_{\text{steps}}} \sum_{s=1}^{N_{\text{steps}}} \phi^{(s)}$
  end for

Chapter 3

Clustering Based SG-MCMC for HMMs

In the implementation of Algorithm 1, one relies on the critical assumption that the gradient estimate in (2.11) is normally distributed about the true gradient such that

$$\widetilde{\nabla U(\theta)} \sim \mathcal{N}(\nabla U(\theta), V(\theta)) \tag{3.1}$$

where $V(\theta)$ is the covariance of the minibatch estimates. When rare states exist within the HMM, this assumption is questionable. Furthermore, even if the estimates are unbiased, one should certainly expect the covariance to be very large given the proclivity of the minibatch samples to exclude rare states. Thus, we focus on designing a subsampling stategy to construct a gradient estimator which i) will avoid bias in the components of the rare latent state(s), ii) have smaller covariance, and iii) be better approximated by a Normal distribution than the original minibatch gradient estimator. While (i) is specific to the case of rare latent state, goals (ii) and (iii) will also improve SG-MCMC in general settings.

To this end, we use stratified sampling instead of minibatches to bias the subsamples to include portions of the time series in the rare state. Again, following the discussion in Section 2.2 we partition the full time series into T/L sequential subseries each of length L which we identify by the index of center point in the subseries, τ_k for $k = \frac{(2k-1)L+1}{2}$ and $k = 1, 2, \ldots, T/L$.

As a preprocessing steps, we first use kmeans++ to cluster the subsequences $\{\mathbf{y}_{\tau_k}\}_{k=1}^{T/L}$ into M clusters, $C_1, \cdots, C_M$. We identify each subseries by its center point so that $C_1, \ldots, C_M$ partition $\{\tau_1, \ldots, \tau_{T/L}\}$. We then draw a subsample $B_m \subset C_m$ for $m = 1, \ldots, M$ independently. Let $n_m = |C_m|$ be the number of subseries in cluster

C_m and $b_m = |B_m|$ be the number of subsamples drawn from C_m. Note that L, M, and $b_1, \ldots, b_M$ are all tuneable parameters which must be specified.

To evaluate (2.9), one could reorder the sum, grouping the sum over subseries together. Therefore, (2.9) can be rewritten as

$$\frac{\partial}{\partial \theta_j} U(\theta) = -\sum_{m=1}^{M} \sum_{\tau \in C_m} \frac{\mathbf{q}^T_{\tau+(L+1)/2} \frac{\partial P(\mathbf{y}_\tau)}{\partial \theta_j} \pi_{\tau-(L+1)/2}}{\mathbf{q}^T_{\tau+(L+1)/2} P(\mathbf{y}_\tau) \pi_{\tau-(L+1)/2}} - \frac{\partial \log p(\theta)}{\partial \theta_j}. \tag{3.2}$$

The inner sum over $\tau \in C_m$ may be approximated using the subsampled indices B_i. Thus, an unbiased estimator of the jth component of the gradient is

$$-\sum_{m=1}^{M} \frac{n_m}{b_m} \sum_{\tau \in B_m} \frac{q^T_{\tau+(L+1)/2} \frac{\partial P(y_\tau)}{\partial \theta_j} \pi_{\tau-(L+1)/2}}{q^T_{\tau+(L+1)/2} P(y_\tau) \pi_{\tau-(L+1)/2}} - \frac{\partial \log p(\theta)}{\partial \theta_j} \tag{3.3}$$

where we rescale each inner sum over B_m by n_m/b_m to match the variance.

The calculation of $\pi_{\tau-(L+1)/2}$ and $q_{\tau+(L+1)/2}$ is still intractable for large T so we follow [MFF17] and make use of the approximation (2.10). Therefore, the stratified estimator $\widehat{\nabla U(\theta)}_i$ is:

$$\frac{\partial}{\partial \theta_j} U(\theta) = -\sum_{m=1}^{M} \frac{n_m}{b_m} \sum_{\tau \in B_m} \frac{\tilde{q}^T_{\tau+(L+1)/2} \frac{\partial P(y_{\tau,L})}{\partial \theta_i} \tilde{\pi}_{\tau-(L+1)/2}}{\tilde{q}^T_{\tau+(L+1)/2} P(y_{\tau,L}) \tilde{\pi}_{\tau-(L+1)/2}} - \frac{\partial \log p(\theta)}{\partial \theta_i} \tag{3.4}$$

Unlike [MFF17], we do not constrain the chosen subsamples within or across clusters to satisfy the gap condition. However, our numerical experiments thus far indicate this estimation procedure indeed meets requirements (i) - (iii). A more detailed analytic investigation is being conducted presently.

Our CSG-MCMC with stratified sampling approach is summarized in Algorithm 2 wherein we have elected to specify a fixed buffer length B explicitly rather than estimate it from the data. This choice was made to simplify the implementation of CSG-MCMC allowing us to focus more directly on improvements in gradient estimation resulting from the stratified subsampling approach.

Algorithm 2 Cluster-enhanced SG-MCMC

Require: Clusters $C_1, \cdots, C_M$ and $b_1, \cdots b_M$
  Select the buffer length B
  Initialize $A^{(0)}$ and $\phi_k^{(0)}$
  for $n = 0, 1, \cdots, N_{iter}$ **do**
    for $s = 1, 2, \cdots, M$ **do**
      $B_s \leftarrow \emptyset$
      for $r = 1, \cdots, b_s$ **do**
        Uniformly sample $i_r \in C_s$ and $B_s \leftarrow B_s \cup \{i_r\}$
      end for
    end for
    Update A and ϕ using $\widehat{\nabla U(\theta)}$ in SG-MCMC as the estimator of $\nabla U(\theta)$
  end for

Chapter 4

Experiments

4.1 Conjugate Emission Distribution

We consider two numerical experiments on synthetic data each with $T = 10^6$ observations and Gaussian emission densities to demonstrate improvements through subseries clustering. For comparison, we report the log predictive density and mean squared loss of the estimated transition matrix as a function of wall-clock run time. Additionally, 10 step predictive intervals and estimated gradient variance are provided.

In the first dataset, hereby referred to the balanced dataset (BD), there are four latent states with transition matrix

$$A_{BD} = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 \\ 0 & 0.9 & 0.1 & 0 \\ 0 & 0 & 0.9 & 0.1 \\ 0.1 & 0 & 0 & 0.9 \end{pmatrix}$$

which has stationary distribution, π_0^{BD}, which is uniform over the states. The Gaussian emission densities have means -6, -3, 0, and 3, and common variance equal to 2.

The second data set is generated from an imbalanced HMM model (ID), with $K = 5$ latent states and transition matrix

$$A_{ID} = \begin{pmatrix} 0.992 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.002 & 0.99 & 0 & 0 & 0 \\ 0.002 & 0 & 0.99 & 0 & 0 \\ 0.002 & 0 & 0 & 0.99 & 0 \\ 0.002 & 0 & 0 & 0 & 0.99 \end{pmatrix}$$

which has an imbalanced stationary distribution $\pi_0^{ID} = (0.5556, 0.1111, 0.1111, 0.1111, 0.1111)^T$ on the hidden states. The Gaussian emission densities have means -20, -10, 0, 10, and 20, and common variance equal to 1. We will use this imbalanced dataset to demonstrate our stratified sampling scheme's ability to capture dynamics. As the emissions have a huge separation and the latent states are not truly rare, one should expect SG-MCMC to work well in this case. Nonetheless, CSG-MCMC outperforms it here too.

In the first case, we subsample the datasets by setting $L = 5$ and the minibatch size $S = 16$. We draw $b_m = 4$ minibatches from each of $M = 4$ clusters. In the second case, we subsample the datasets by setting $L = 5$, $M = 5$, and draw 1 sample from each cluster for a total sample size $S = 5$. To cluster the data we view the subseries as vector in $\mathbb{R}^L$ and use the Euclidean metric in kmeans++.

We assign improper (constant) priors for all parameters so that $\nabla \log p(\theta) = 0$. In this case, the parameters are the emission means and variances and A. The simplex constraints on the columns of A are attained by projecting back to the simplex after each iteration.

We plot the 10-step-ahead log predictive density of 2000 alternative time observations and 95% predictive intervals to evaluate the performance of both algorithms. In the BD dataset, CSG-MCMC outperforms SG-MCMC in terms of log predictive density (Figure 4.1). The clustering-based method requires less runtime to attain the same predictive density, which indicates that our method has a faster convergence speed. Moreover, in the long run, its log predictive density is still higher than that of the other algorithm, which suggests that it provides a more accurate approximation to the posterior.

Concerning predictive intervals, the non-clustering predictive intervals fail to include the first observation most of the time, and they are wider than the non-clustering intervals (Figure 4.2). CSG-MCMC outperforms SG-MCMC due to the greatly decreased variance in the gradient estimates $\nabla U(\theta)$ in every iteration (See Figure 4.3).

In the study of the ID dataset, we use $||A - A_{true}||_2$ to evaluate the performance of our clustering-based algorithm with respect to capturing rare dynamics. The convergence performance of A indicates that the clustering-based algorithm is approximately 6 times faster. Furthermore, the long time behavior of both algorithms indicates that the clustering-based algorithm provides a more accurate estimate of the posterior. The reason why the clustering-based algorithm outperforms the non-clustering algorithm is that the stratified sampling scheme guarantees rare dynamics are selected to estimate the gradient in every iteration, while the non-clustering algorithm will likely fail to incorporate the information from rare dynamics.

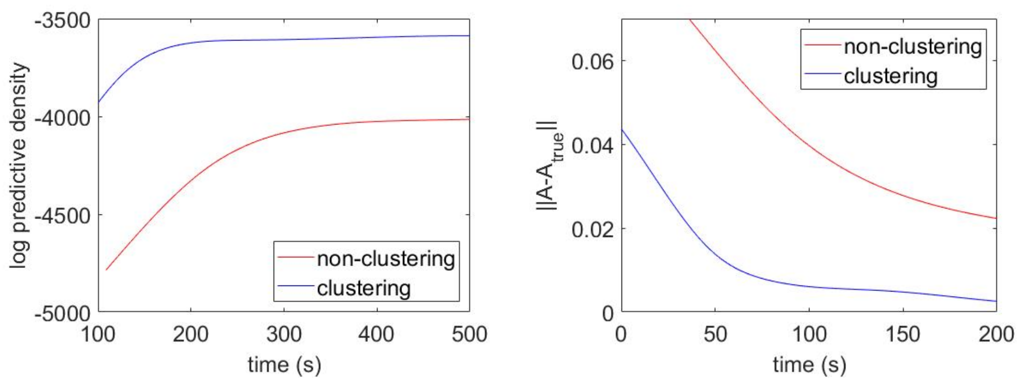

Figure 4.1: The log predictive density of 2000 alternative observations versus time in the BD dataset. $||A - A_{true}||_2$ versus time in the ID dataset.

4.2 Non-conjugate Emission Distribution

As a second experiment to demonstrate the advantages of the CSG-MCMC, we study the convergence speed and accuracy when a non-conjugate emission probability is specified. In this case, we simulate a dataset with $T = 3 \times 10^5$ observations and

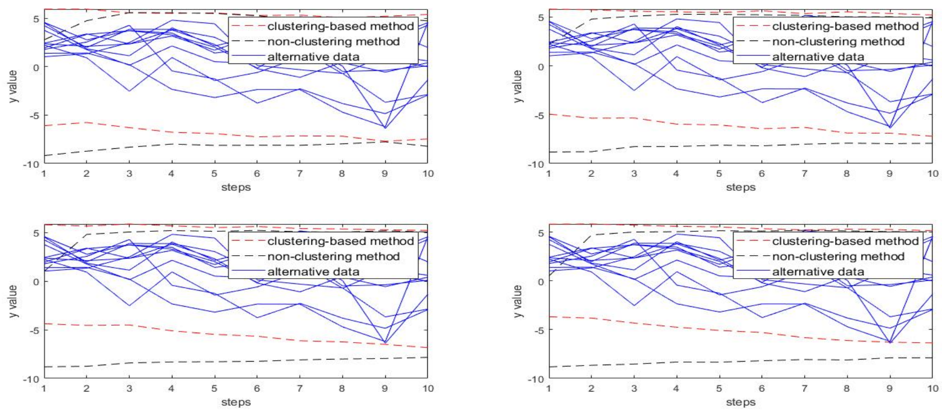

Figure 4.2: The 95% predictive intervals of both methods and 10 realizations of the generated alternative data.

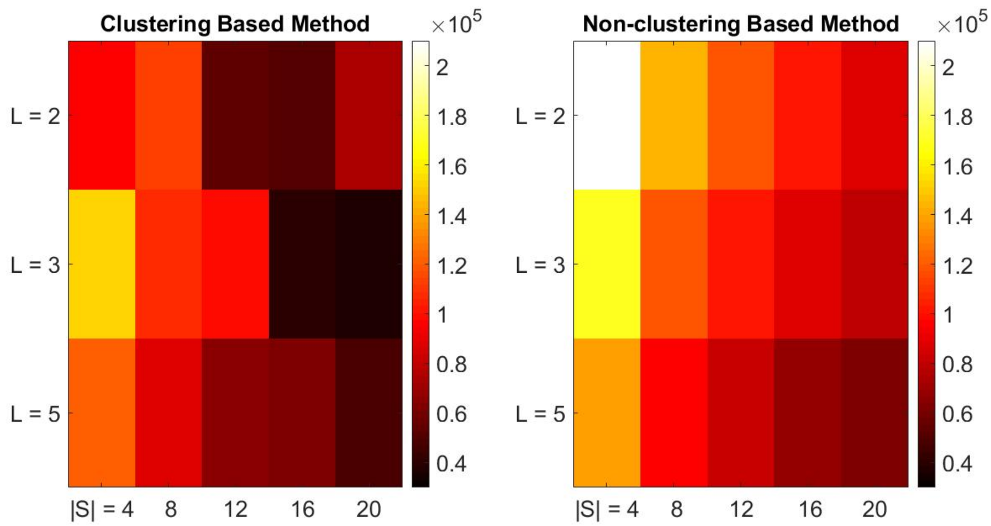

Figure 4.3: The Monte Carlo estimates of the variances of estimated $\nabla U(\theta)$ with different minibatch size $|S|$ and L at a point in the parameter space.

$K = 2$ latent states with the transition matrix

$$A = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

The stationary distribution is uniform over the two states so there are no rare states. The emission distribution of the first state is $\text{Ber}(0.9)$, while the emission distribution of second is $\text{Ber}(0.1)$.

We subsample the time series by setting $L = 5$ and the total subsample size to $S = 4$ comprised of $b_1 = b_2 = 2$ samples from each of $M = 2$ clusters. We use uniform priors for the success probabilities in the Bernoulli emissions and improper priors for A.

Similar to Section 4.1, we plot the log predictive density of 2000 alternative observations and use $||A - A_{true}||_2$ to evaluate the estimation performance of both algorithms (Figure 4.4).

The log predictive density of alternative data shows CSG-MCMC has both a higher convergence speed and better accuracy. Furthermore, the convergence performance of the transition parameter A demonstrates that the clustering-based method provides an accurate estimate of A in an approximately 10 times higher speed.

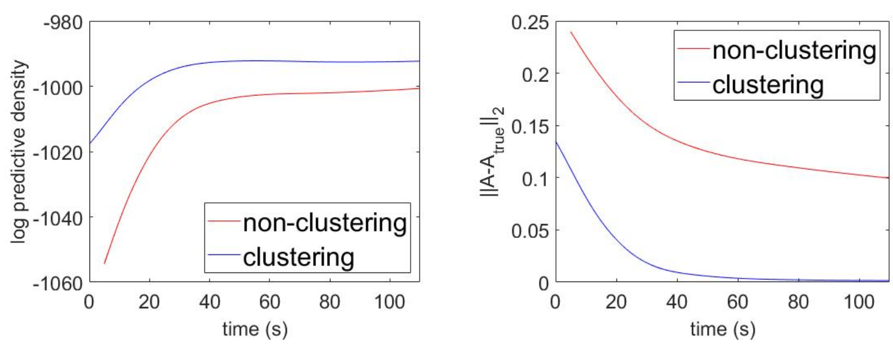

Figure 4.4: The log predictive density of 2000 alternative observations versus time in the non-conjugate model. $||A - A_{true}||_2$ versus time.

4.3 Real Data Analysis

The discovery and prediction of rare states is of particular importance in heart rate data wherein a patient of interest may exhibit normal behavior most of the time with intermittent moments with a dangerously rapid or slow pulse. The increasing prevalence of wearable technology has made it possible to attain second-to-second heart rate data over many days resulting in massively long time series. In analyzing

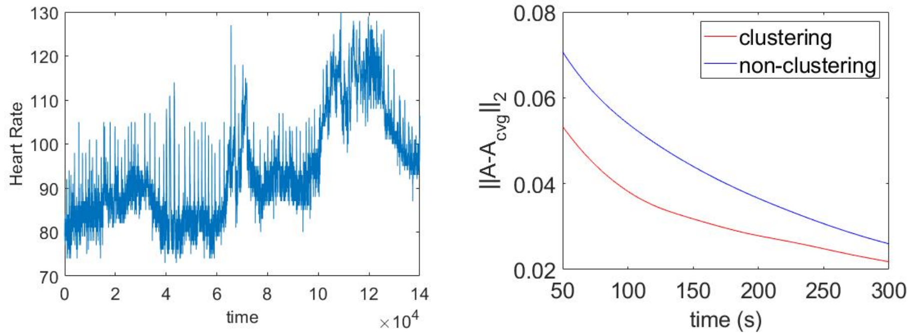

Figure 4.5: The heart rate of a single patient (Left). The plot of $||A - A_{cvg}||_2$ versus time of both algorithms (Right).

such time series, naive subsampling is likely to miss rare dynamics, thereby hindering detection, let alone inference, of such rare states.

In this section, we analyze heart rate data from the MIMIC-III database using both CSG-MCMC and SG-MCMC in which heart rate is reported at approximately one second intervals. We restrict our focus to the 84 largest time series. An example of one such series is shown in Figure 4.5.

We build a model with $K = 5$ hidden states and Gaussian emission distribution in each state. Due to the computational budget, we randomly select $T = 1804987$ in total and run the algorithms for 1500 iterations each. Since recognizing the states of patients are crucial, we mainly focus on how both algorithms perform with regard to recognizing dynamics. We use $||A - A_{cvg}||_2$, where A_{cvg} is the average of the transition matrix over the last 200 iterations of the algorithm, as our metric of performance.

We plot the comparison of both algorithms in Figure 4.5. The plot shows that the transition parameter A achieves stationarity faster in CSG-MCMC. Moreover, when the runtime is long, the accuracy of the clustering-based method still dominates that of the other. Thus, the clustering-based method provides a more accurate and faster estimate of the transition parameter A.

Chapter 5

Discussion

The numerical experiments in the previous section indicate clear improvements in CSG-MCMC but also elicit a number of questions worth addressing. We are currently interested in pursuing two primary directions.

First, the results of Figure 4.3 demonstrate a reduction in variance of the CSG-MCMC in gradient estimation. However, a more thorough review is necessary to specifically address the issues of normality, bias, and variance of the clustering based estimation procedure. Rigorous results would then allow for a careful investigation of the optimal situations in which CSG-MCMC can provide substantial gains in addition to detailed bounds on bias in posterior estimates and potential avenues for improvement.

Additionally, there are number of different approaches one could use in the pre-processing clustering step. Throughout the present work, we clustered subseries by viewing them as L-dimension vectors and using kmeans++ with the Euclidean metric. This approach has clear limitations with regards to clustering subseries with rare extreme data. For example, consider the original motivating example from Section 2.1. The spikes associated with the rare state could occur at any one of the L points within a given subseries. As a result, there will be $L + 1$ clusters in $\mathbb{R}^L$. Most of the subseries will cluster near the origin with L additional clusters near each of the canoncial vectors $\mathbf{e}_i$, $i = 1, \ldots, L$. Clustering based on Euclidean distance is unlikely to group data near $\mathbf{e}_i$ and $\mathbf{e}_j$, $i \neq j$ potentially absorbing rare event subseries into clusters comprised of typical dynamics when $M < L + 1$. As a result, this will eliminate the biased subsampling CSG-MCMC was intended to address.

Alternatively, one could attempt an alignment of the cluster vectors in a myriad of manners such as Procrustes' analysis or more simply sorting the data points within each subseries prior to embedding in $\mathbb{R}^L$. While these two approaches add computation to the preprocessing step them may greatly improve the grouping of rare event data into a single cluster. Other strategies include clustering based on one or more summary statistics of the subseries. This would shorten the clustering step if the number of summary statistics is less than L.

Finally, both the gradient estimation and clustering will depend on details of the HMM, perhaps in drastic and in surprising ways. Thus, additional numerical experimentation is needed to identify setting specific choices for the L, S, M and clustering methodology within CSG-MCMC.

Chapter 6

Conclusion

We propose a new MCMC algorithm to approximately sample from the posterior distribution of hidden Markov models. We demonstrate very substantial gains in predictive and inferential accuracy on real and synthetic examples.

Bibliography

[Bea03] Matthew James Beal. Variational algorithms for approximate Bayesian inference. *Ph.D. Thesis*, 2003.

[Bis06] Christopher Bishop. *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer-Verlag, New York, 2006.

[CB17] Trevor Campbell and Tamara Broderick. Automated Scalable Bayesian Inference via Hilbert Coresets. *arXiv:1710.05053 [cs, stat]*, October 2017.

[CHSP04] M. T. Chan, A. Hoogs, J. Schmiederer, and M. Petersen. Detecting rare events in video using semantic primitives with HMM. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, pages 150–154 Vol.4, August 2004.

[FXLF14] Nick Foti, Jason Xu, Dillon Laird, and Emily Fox. Stochastic variational inference for hidden Markov models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3599–3607. Curran Associates, Inc., 2014.

[FZ17] Tianfan Fu and Zhihua Zhang. CPSG-MCMC: Clustering-Based Preprocessing method for Stochastic Gradient MCMC. In *Artificial Intelligence and Statistics*, pages 841–850, April 2017.

[GKQ+18] David Gunawan, Robert Kohn, Matias Quiroz, Khue-Dung Dang, and Minh-Ngoc Tran. Subsampling Sequential Monte Carlo for Static Bayesian Models. *arXiv:1805.03317 [stat]*, May 2018.

[HAJ90] Xuedong D Huang, Yasuo Ariki, and Mervyn A Jack. Hidden Markov models for speech recognition. 1990.

[HCB16] Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for Scalable Bayesian Logistic Regression. In *30th Conference on Neural Information Processing Systems*, page 9, Barcelona, Spain, 2016.

[HN05] M. R. Hassan and B. Nath. Stock market forecasting using hidden Markov model: A new approach. In *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pages 192–196, September 2005.

[HSS15] Michael C Hughes, William T Stephenson, and Erik Sudderth. Scalable Adaptation of State Complexity for Nonparametric Hidden Markov Models. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and

R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1198–1206. Curran Associates, Inc., 2015.

[HW13] Eric C. Hall and Rebecca M. Willett. Online Optimization in Dynamic Environments. *arXiv:1307.5944 [cs, math, stat]*, July 2013.

[JBP+02] S. V. Jones, S. Brown, T. Parker, V. Stubblefield, E. Kössler, C. Simmons, and J. Wittner. Anomalous indeterminate amorphous transitory gradations observed in certain specific nongeneralizable locii attendant to unique fleeting unpredictable phenomena. *Journal of Nonspecific Research*, 88:665–703, 2002.

[Joh14] Matthew James Johnson. *Bayesian Time Series Models and Scalable Inference*. Thesis, Massachusetts Institute of Technology, 2014.

[JPS+16] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, May 2016.

[JW14] M.J. Johnson and A.S. Willsky. Stochastic variational inference for Bayesian time series models. In *31st International Conference on Machine Learning, ICML 2014*, pages 3872–3880, 2014.

[JW16] Xin Jiang and Rebecca Willett. Online Data Thinning via Multi-Subspace Tracking. *arXiv:1609.03544 [cs, stat]*, September 2016.

[MCF15] Yi-An Ma, Tianqi Chen, and Emily B. Fox. A Complete Recipe for Stochastic Gradient MCMC. *Advances in Neural Information Processing Systems*, 28:2899–2907, 2015.

[MFF17] Yi-An Ma, Nicholas J. Foti, and Emily B. Fox. Stochastic Gradient MCMC Methods for Hidden Markov Models. 2017.

[Naw28] B. Nawahi. Acoustic production in liquid filled ceramic environments. *J. Chem. Phys.*, 108:9893–9904, 1928.

[OMSH03] D. Ourston, S. Matzner, W. Stump, and B. Hopkins. Applications of hidden Markov models to detecting multi-stage network attacks. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of The*, pages 10 pp.–, January 2003.

[PCX16] Anastasios Petropoulos, Sotirios P. Chatzis, and Stylianos Xanthopoulos. A novel corporate credit rating system based on Student's-t hidden Markov models. *Expert Systems with Applications*, 53:87–105, July 2016.

[QKVT18] Matias Quiroz, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. Speeding Up MCMC by Efficient Data Subsampling. *Journal of the American Statistical Association*, 0(0):1–13, March 2018.

[QTVK18] Matias Quiroz, Minh-Ngoc Tran, Mattias Villani, and Robert Kohn. Speeding up MCMC by Delayed Acceptance and Data Subsampling. *Journal of Computational and Graphical Statistics*, 27(1):12–22, January 2018.

[San05] F. Sanford. Oxidized ferrous materials. *Journal of Junk*, 5(4):324–345, 2005.

[SBB+16] Steven L. Scott, Alexander W. Blocker, Fernando V. Bonassi, Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, April 2016.

[Sco02] Steven L. Scott. Bayesian Methods for Hidden Markov Models. *Journal of the American Statistical Association*, 97(457):337–351, March 2002.

[SD18] Abhra Sarkar and David B. Dunson. Bayesian Nonparametric Higher Order Hidden Markov Models. *arXiv:1805.12201 [stat]*, May 2018.

[SH70] M. Smith and I. A. Hall. *Handbook of Interstellar Travel*. Dover, New York, 7th edition, 1970.

[SZG+11] Johannes Stigler, Fabian Ziegler, Anja Gieseke, J. Christof M. Gebhardt, and Matthias Rief. The Complex Folding Network of Single Calmodulin Molecules. *Science*, 334(6055):512–516, October 2011.

[ZGBM05] Dong Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Semi-supervised adapted HMMs for unusual event detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 611–618 vol. 1, June 2005.